Tornadoes cost - R notebook

January 29, 2017

In [6]:
```
library("AzureML")
ws <- workspace()
dat <- download.datasets(ws, "All_tornadoes.csv")
```

In [7]:
```
head(dat)
```

NA.	Year	Month	Day	Date	Time	Timezone	State	State_FIPS	State_Number	...	Ler
1	1950	1	3	3/1/1950	11:00	3	MO	29	1	...	9.5
1	1950	1	3	3/1/1950	11:00	3	MO	29	1	...	6.2
1	1950	1	3	3/1/1950	11:10	3	IL	17	1	...	3.3
2	1950	1	3	3/1/1950	11:55	3	IL	17	2	...	3.6
3	1950	1	3	3/1/1950	16:00	3	OH	39	1	...	0.1
4	1950	1	13	13/1/1950	5:25	3	AR	5	1	...	0.6

0.1 Replace month values with names

In [8]:
```
dat$Month[dat$Month == 1] <- "January"
dat$Month[dat$Month == 2] <- "February"
dat$Month[dat$Month == 3] <- "March"
dat$Month[dat$Month == 4] <- "April"
dat$Month[dat$Month == 5] <- "May"
dat$Month[dat$Month == 6] <- "June"
dat$Month[dat$Month == 7] <- "July"
dat$Month[dat$Month == 8] <- "August"
dat$Month[dat$Month == 9] <- "September"
dat$Month[dat$Month == 10] <- "Octomber"
dat$Month[dat$Month == 11] <- "November"
dat$Month[dat$Month == 12] <- "December"
```

In [9]:
```
head(dat)
```

NA.	Year	Month	Day	Date	Time	Timezone	State	State_FIPS	State_Number	...	Le
1	1950	January	3	3/1/1950	11:00	3	MO	29	1	...	9.5
1	1950	January	3	3/1/1950	11:00	3	MO	29	1	...	6.2
1	1950	January	3	3/1/1950	11:10	3	IL	17	1	...	3.3
2	1950	January	3	3/1/1950	11:55	3	IL	17	2	...	3.6
3	1950	January	3	3/1/1950	16:00	3	OH	39	1	...	0.1
4	1950	January	13	13/1/1950	5:25	3	AR	5	1	...	0.6

0.2 Estimated property loss information

Prior to 1996 this is a categoration of tornado damage by dollar amount: - 0 or blank-unknown - 1 < $50 - 2=$50-$500 - 3=$500-$5,000 - 4=$5,000-$50,000 - 5=$50,000-$500,000 - 6=$500,000-$5,000,000 - 7=$5,000,000-$50,000,000 - 8=$50,000,000-$500,000,000 - 9> $500,000,000

When summing for state total use sn=1, not sg-1.

From 1996, this tornado property damage is in million of dollars. Entry of 0 does not mean $0.

```
In [10]: # Before 1996
         dat$Losses[dat$Losses == 0 ] <- NA
         dat$Losses[dat$Losses == 1 & dat$Year < 1996] <- "<$50"
         dat$Losses[dat$Losses == 2 & dat$Year < 1996] <- "$50-$500"
         dat$Losses[dat$Losses == 3 & dat$Year < 1996] <- "$500-$5,000"
         dat$Losses[dat$Losses == 4  & dat$Year < 1996] <- "$5,000-$50,000"
         dat$Losses[dat$Losses == 5 & dat$Year < 1996] <- "$50,000-$500,000"
         dat$Losses[dat$Losses == 6 & dat$Year < 1996] <- "$500,000-$5,000,000"
         dat$Losses[dat$Losses == 7 & dat$Year < 1996] <- "$5,000,000-$50,000,000"
         dat$Losses[dat$Losses == 8  & dat$Year < 1996] <- "$50,000,000-$500,000,00
         dat$Losses[dat$Losses == 9 & dat$Year < 1996] <- ">$500,000,000"

         # After 1996
         dat$Losses[dat$Losses >= 0 & dat$Losses < 0.0005 & dat$Year >= 1996] <- "<
         dat$Losses[dat$Losses >= 0.00005 & dat$Losses < 0.0005 & dat$Year >= 1996]
         dat$Losses[dat$Losses >= 0.0005 & dat$Losses < 0.005 & dat$Year >= 1996] <
         dat$Losses[dat$Losses >= 0.005 & dat$Losses < 0.05 & dat$Year >= 1996] <-
         dat$Losses[dat$Losses >= 0.05 & dat$Losses < 0.5 & dat$Year >= 1996] <- "$
         dat$Losses[dat$Losses >= 0.5 & dat$Losses < 5 & dat$Year >= 1996] <- "$500
         dat$Losses[dat$Losses >= 5 & dat$Losses < 50 & dat$Year >= 1996] <- "$5,00
         dat$Losses[dat$Losses >= 50 & dat$Losses < 500 & dat$Year >= 1996] <- "$50
         dat$Losses[dat$Losses >= 500 & dat$Year >= 1996] <- ">$500,000,000"
```

```
In [11]: newdata <- subset(dat,select=c(Date, State, Losses))
         head(newdata)
```

Date	State	Losses
3/1/1950	MO	$500,000-$5,000,000
3/1/1950	MO	$500,000-$5,000,000
3/1/1950	IL	$50,000-$500,000
3/1/1950	IL	$50,000-$500,000
3/1/1950	OH	$5,000-$50,000
13/1/1950	AR	$500-$5,000

0.3 Set Crop-loss NA

```
In [12]: dat$'Crop.loss'[dat$'Crop.loss' == 0] <- NA
```

0.4 Set Length(miles) NA

```
In [14]: dat$'Length.miles.'[dat$'Length.miles.' == 0] <- NA
```

0.5 Set Width(yards) NA

```
In [15]: dat$'Width.yards.'[dat$'Width.yards.' == 0] <- NA
```

```
In [16]: newdata <- subset(dat,select=c(Date, State, Crop.loss,Length.miles. , Widt
         head(newdata)
```

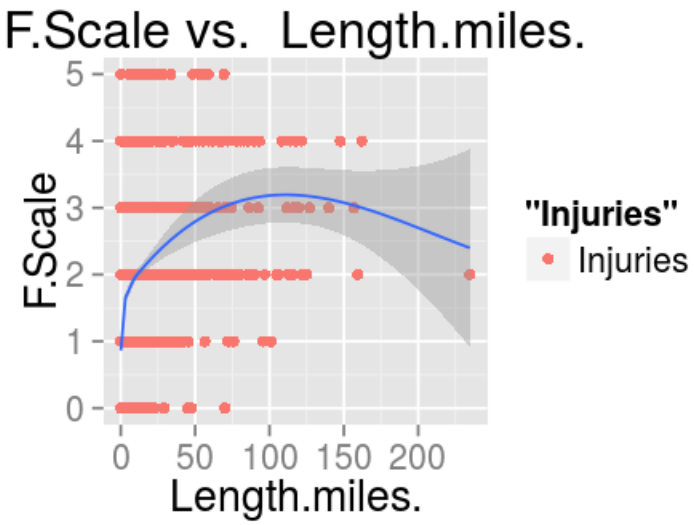
Date	State	Crop.loss	Length.miles.	Width.yards.
3/1/1950	MO	NA	9.5	150
3/1/1950	MO	NA	6.2	150
3/1/1950	IL	NA	3.3	100
3/1/1950	IL	NA	3.6	130
3/1/1950	OH	NA	0.1	10
13/1/1950	AR	NA	0.6	17

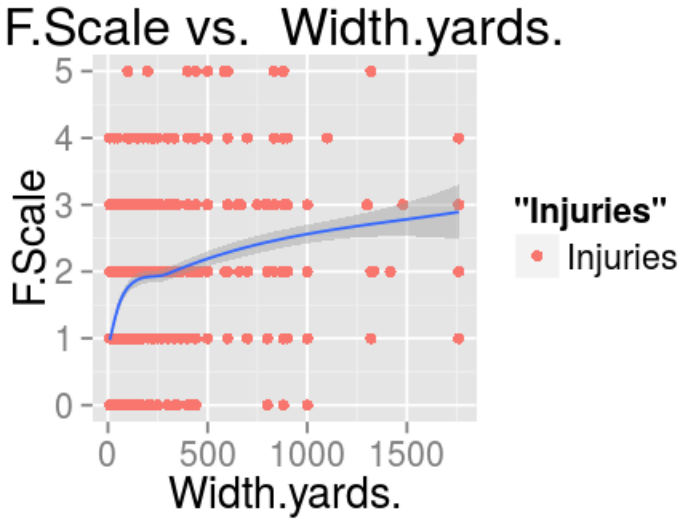
```
In [17]: library(ggplot2)
```

```
In [47]: ## Use ggplot2 to create conditioned scatter plots
         numCols <- c( 'Length.miles.','Width.yards.' )
         fscale.scatter <- function(df, cols){
           require(ggplot2)
           for(col in cols){
             p1 <- ggplot(df, aes_string(x = col, y = "F.Scale")) +
                     geom_point(aes( color = 'Injuries')) +
                     geom_smooth(method = "loess") +
                     ggtitle(paste('F.Scale vs. ', col)) +
                     theme(text = element_text(size=16))
             print(p1)
           }
         }
```

```
In [48]: fscale.scatter(dat[1:4000,], numCols)
```

```
Warning message:
In loop_apply(n, do.ply): Removed 25 rows containing missing values (stat_smooth).W
In loop_apply(n, do.ply): Removed 25 rows containing missing values (geom_point).Wa
In loop_apply(n, do.ply): Removed 25 rows containing missing values (stat_smooth).W
In loop_apply(n, do.ply): Removed 25 rows containing missing values (geom_point).
```

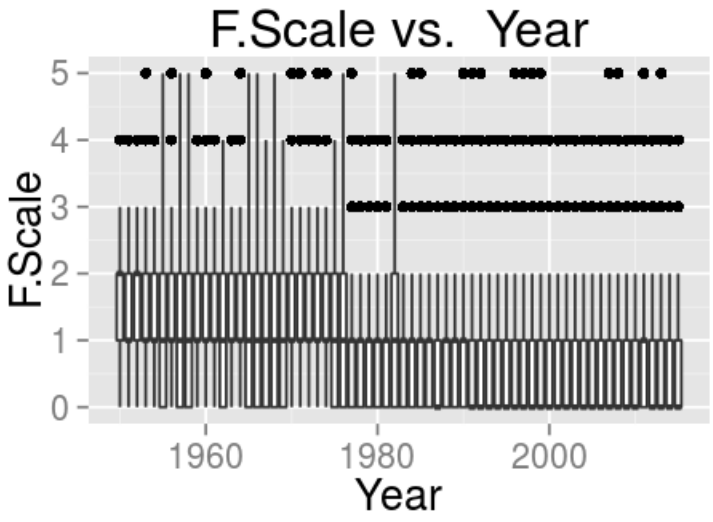


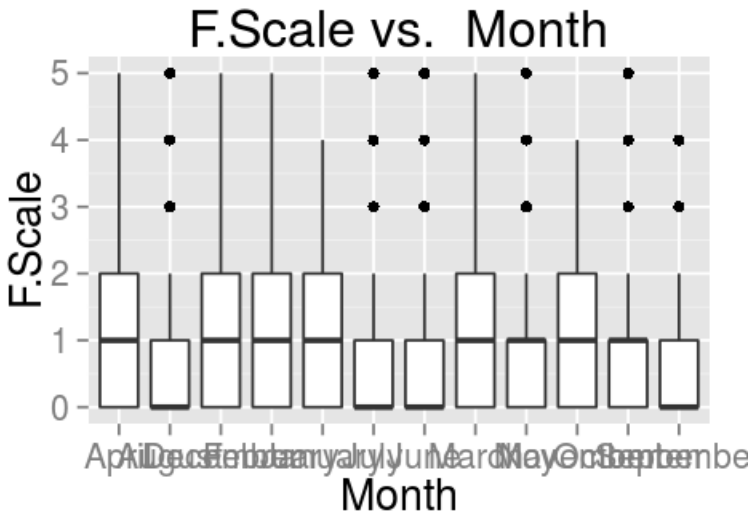


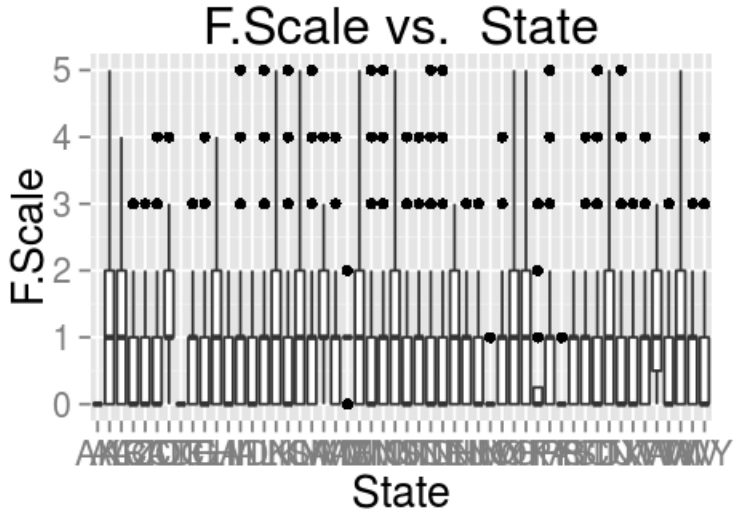

```
In [58]: catCols <- c('Year', 'Month', 'State','Losses')
         tornadoes.box <- function(df, cols){
           require(ggplot2)
           for(col in cols){
             p1 <- ggplot(df, aes_string(x = col, y = 'F.Scale', group = col)) +
                     geom_boxplot()+
                     ggtitle(paste('F.Scale vs. ', col)) +
                     theme(text = element_text(size=16))
```

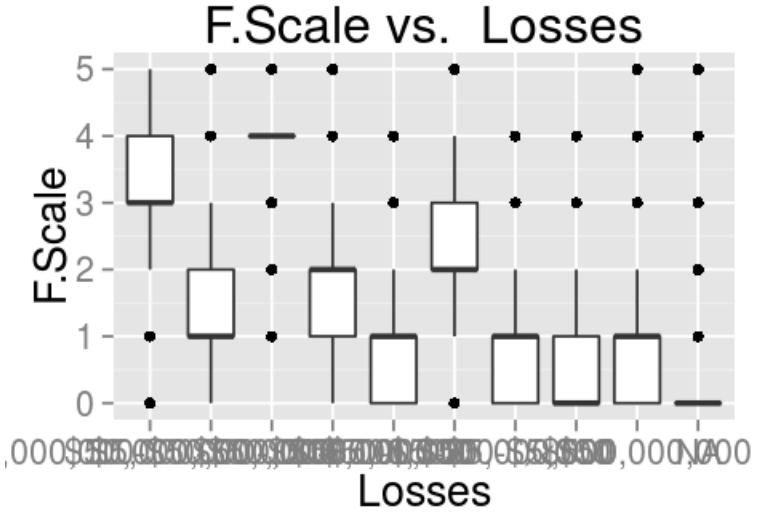
```
            print(p1)
          }
        }
```

In [59]: `tornadoes.box(dat, catCols)`






F.Scale vs. State
F.Scale
5
4
3
2
1
0
State


F.Scale vs. Losses
F.Scale
5
4
3
2
1
0
Losses